Chapter 20 Bid-ask Spreads, Commissions and Other Costs

Thanos Verousis
Lecturer (Assistant Professor), Bangor Business School, Bangor University, UK.

KEYWORDS: order processing cost; adverse selection; inventory cost; high frequency trading; electronic trading; tick size.

ABSTRACT

This chapter examines trading costs that are associated with the "costs of exchanging ownership titles". Costs are categorized as (i) commission charges which are determined by the exchange and (ii) cost components of the bid-ask spread that are determined by market participants. There are three main components of the bid-ask spread. The order processing cost which is associated with the cost of providing liquidity, the inventory cost that is due to short-term order imbalances and the adverse selection cost that refers to the cost of trading with informed traders. Spreads and commission charges are currently at very low levels in developed markets which have led to a great expansion in algorithm trading and trading volume. Trading costs for emerging markets appear to be considerably higher than for the more developed markets. Market capitalisation and liquidity differences explain some of the variability in trading costs in exchanges around the world. A second element of variability is attributed to differences in market structures.

INTRODUCTION

In the market microstructure literature, the majority of research is concentrated on or related to the existence of transaction costs, and in particular to the origins and determinants of the bid-ask spread. The seminal work on bid-ask spread costs is by Demsetz (1968) hence it is only fair to start this chapter with the original definition of transaction costs. Demsetz (1968) defined transaction costs as the "cost of exchanging ownership titles", which is further reduced to the bid-ask spread and commission charges when all the costs of issuing the titles and the costs of being informed about the titles have been incurred.

Recent evidence suggests that commission charges are currently at historically all-time-low levels, which is mainly an outcome of two major changes in financial markets: the move to decimal pricing and market consolidation have led to increases in the economies of scale in the cost of handling orders (see Stoll, 2006). A study on commission costs by Jones (2002) shows that commission costs on the New York Stock Exchange (NYSE) amount at about 0.12% of the amount transacted, down from about 24% in 1994. Historically, until 1974, commission charges in the U.S. have been linear in nature. That is, commission charges increased proportionately with trade size, however, the deregulation of commission charges of 1974 led to economies of scale in commission costs. Similarly, trading costs that are associated with the existence of the bid-ask spread have declined substantially (see Chung et al., 2004 and Stoll, 2000).

The decline in trading costs has led to a substantial increase in trading volume but most importantly has opened the way for the widespread implementation of algorithm trading, which refers to computerised trade strategies that rely on heavy volume to extract marginal profits from a very large number of trades. In 2009, High Frequency Trading (HFT) firms accounted for approximately 2% of the total number of trading firms, however they were responsible for almost 75% of the total trading volume in the U.S., approximately a threefold increase from 2006 (Iati, 2009).

A question that naturally arises regarding aggregate trading costs is, why does the bid-ask spread exist? Early studies on this field have shown that spreads reflect the price of "immediacy" (Demsetz, 1968). Traders that wish to sell (buy) an asset now have no guarantee that a buyer (seller) will be available in the market, hence buyers (sellers) will only agree to trade if they are compensated for the immediate provision of their services. In this context, the spread reflects order processing costs. A second spread component arises when factoring in the total inventory costs that are associated with the fact that trade imbalances occur over a trading period (see Smidt, 1971). Finally, a third cost component arises when incorporating the fact that market participants have asymmetric information, hence traders that choose to trade with informed traders effectively provide liquidity at a cost. In order to recover these costs, market participants charge the spread to liquidity traders (see Bagehot, 1971).

The above decomposition of spreads to order processing, inventory and adverse selection costs, has explained trading cost variability in relation to market participants. A second level of variability is further explained when looking at differences in market structures. Studies have shown that even when controlling for firm-level and market liquidity differences, trading cost differences across exchanges around the world can be attributed to market structure differences (see Jain, 2003). Studies have shown that electronic trading systems enhance liquidity by reducing trading costs, however, floor-based trading systems have been shown to provide smaller costs for large institutional trades. Similarly, costs fell substantially with the introduction of decimal pricing in the US (see Bessembinder, 2003).

WHY DOES THE BID-ASK SPREAD EXIST?

Order Processing Costs

Demsetz (1968) refers to the bid-ask spread as the “price of immediacy” and traces its origins to the problem of simultaneity in buying and selling securities. In particular, a trader wishing to buy an asset has no guarantee that a seller will be readily available in the market to provide this service. A trader who is able to provide this service to the buyer will only agree to sell the asset at a price greater than the price the trader has paid to buy this asset. Also, a seller who wishes to sell an asset immediately will agree to wait and sell at a future date only if the trader is compensated for waiting to trade at a future date. Similarly, a trader wishing to sell an asset has no guarantee that there will be a buyer available in the market, hence the trader who is available to provide this service immediately will have to be compensated for providing this service and a buyer who stands ready to buy an asset will have to be compensated for the cost of waiting to buy at a future date. In this respect, the bid-ask spread exists as the “price of immediacy”. If buy and sell orders arrived simultaneously the market would clear at the equilibrium price and a spread would not exist.

That is, traders conduct two transactions, one at the ask and one at the bid price, instead of one if the simultaneity problem did not exist.

In a limit order market, the price of immediacy is reflected in the limit order book: traders wishing to trade at a later time will post limit orders, traders wishing to trade immediately will submit market orders. In this respect, ask and bid prices will not carry any extra charges apart from commission costs. In a dealer market, dealers (or specialists) will not only facilitate the matching of buyers with sellers (assuming a small commission) but most importantly have an obligation to "trade against the flow", that is a dealer also trades for his own account hence the spread also reflects the dealer's income. The original paper by Demsetz (1968) shows that 40% of the total transaction cost of stocks trading at the NYSE reflects the spread component and 60% of the total transaction costs is commission charges.

Inventory Costs

To the extent that bid and ask prices reflect the price of immediacy, the spread primarily consists of order processing (or immediacy) and commission costs. However, subsequent studies have shown that dealers are not to be seen as risk neutral as far as their inventory control is concerned. Smidt (1971) argued that dealers' requirement to provide liquidity is often in disagreement with their goal of profit maximisation, which is conditional on the assumption that dealers will not go bankrupt or fail. Garman (1976) rationalised this idea further. In Garman's (1976) model, the dealers' optimisation problem is modelled as a series of independent buy and sell orders that come at irregular times at specified arrival probabilities. Under these conditions, dealers will not only adjust spreads but also bid and ask prices in order to make sure that the probability of bankruptcy does not equal one. For example, dealers will decrease their bid price when they do not want to increase their inventory further (see also Madhavan, 2000). Several papers also emphasize the notion that the dealer's inventory control problem causes a deterioration of dealer's optimal inventory positions hence spreads exist in order to make up for the anticipated losses incurred from assuming unnecessary risk (see Stoll, 1978; Amihud and Mendelson, 1980). A second deviation from Garman's (1978) risk neutral model comes in Ho and Stoll (1981) and in O'Hara and Oldfield (1986) amongst others. In Ho and Stoll (1981), a dealer's expected utility of terminal wealth is maximised by adjusting bid and ask prices through time, hence bid and ask prices are a function of trade size, stock price volatility, dealer's time horizon and the risk aversion coefficient. O'Hara and Oldfield (1986) show that inventories affect both the size and the placement of dealers' spreads and also that risk averse dealers will set different spreads to risk neutral dealers.

Adverse Selection Costs

Alongside these two streams of literature, a third class of models emerged which was based on the notion that truly informed traders may exist in the market in the sense that there are three types of traders, liquidity traders, dealers and informed traders and that both classes of liquidity and dealer traders possess less information than the informed traders (Bagehot, 1971; Logue, 1975; Jaffee and Winkler, 1976). These models conclude that even under competitive dealer markets with risk neutral dealers, spreads will exist reflecting an adverse selection cost component of the spread. In Copeland and Galai's (1983) adverse selection cost model, the dealer sets spreads that maximize his profits. However, with the inclusion of liquidity and informed traders, if the spread is too wide, the dealer faces losing profits from limited trading with liquidity traders and if the spread is too narrow, the dealer faces the risk of losses from the informed traders. The model's main prediction is consistent with the adverse selection theory of the spread: dealers will always set an ask price as a mark-up from the true equilibrium price and a bid price as a mark-down from the equilibrium price.

Glosten and Milgrom (1985) further improved this model by addressing its static nature, whereby transaction prices are informative, which results in declining spread patterns. Nevertheless, the main predictions of the adverse selection theory remain, namely that the spread is a function of the adverse selection cost and that the spread would still exist even with zero-profit, risk-neutral dealers.

MEASURES OF BID-ASK SPREAD AND OTHER TRADING COSTS

Commission Charges

The above section establishes the three main components of the trading costs; the order processing cost, the inventory cost and the adverse selection cost. A fourth component refers to commission charges, however, one important difference between the trading costs that are reflected by the existence of the bid-ask spread and commission costs is the fact that spread costs are determined by market participants acting independently whereas commission charges are determined by the exchange (Demsetz, 1968). Of course, the minimum spread cost is also a function of the minimum price increment, if such a rule applies at the exchange.

Historically, commission charges have been very volatile. Jones (2002) tracks the evolution of commission costs for NYSE stocks over the period 1925-2000. The author shows that, between 1925 and 1975, when all commissions were heavily regulated and no commission discounts were allowed, one-way proportional commission charges rose from 0.27 % in 1925 to 0.90 % in 1974. Since the exchange deregulation of commission costs, spreads continued to decline by about half every seven or eight years (see Jones, 2002).

Stoll (2006) shows that the technological advances with the increased use of electronic trading have substantially reduced the cost of handling orders by broker-dealers even further, hence commission charges are at very low levels. Stoll (2006) reports that the average round-trip commission as percentage of public dollar volume for shares trading in the U.S. has decreased from 1.17% in 1980 to 0.21% in 2001. Also, Jones and Lipson (2001) show that one-way institutional commissions for U.S. stocks listed on the NYSE amounted to about 0.12% of the amount transacted, down from 0.24% in 1994 (see Keim and Madhavan, 1997 and Jones, 2002).

Measuring Bid-Ask Spreads

Assuming that commission charges remain constant over a short time period, the variability in trading costs will reflect the existence of the bid-ask spread and the deviation that the spread reflects between the “true” equilibrium price and the bid and ask prices. This difference between the “true” price and the trading price is referred to as the trade execution cost. The simplest measure of implied execution costs is the quoted spread, which is usually denoted in percentage basis points. The quoted spread of a stock is calculated as the ratio of quoted spread (ask – bid) over the quote midpoint. Quoted spreads are calculated on an intraday basis, usually at fixed time intervals, and several studies have shown an intraday U-shaped pattern in the quoted spread (see Chan et al., 1995). However, quoted spreads are only implicit measures of execution costs, as they do not refer to the actual traded price. A second, explicit measure of trading costs is the effective bid-ask spread, which is estimated as the ratio of the absolute difference between the traded price and the quote midpoint over the quote midpoint (see Huang and Stoll, 2001). In markets where trade negotiations are allowed, the effective spread also reflects trade improvements as traders are permitted to trade “inside” the quotes, hence the effective spread tends to be lower than the quoted spread. A second measure of implicit trading costs is the price impact which measures adverse selection costs, that is the costs of trading with an informed trader (see Stoll, 2006; Bessembinder, 1999). Price impact costs are measured as the percentage difference between the midquote that prevailed at the time of the trade and a future midquote. A final measure of transaction costs is the realised spread which is estimated as the difference between the effective spread and price impact. As the realised spread is net of the price impact, it reflects trading costs net of any losses to informed traders.

Stoll (2000) shows that the average quoted half-spread for stocks trading on the NYSE is 7.87 cents per share and the average effective half-spread is 5.58. Nevertheless, Chung et al. (2004) show that effective (quoted) spreads on the NYSE fell by approximately 40 (36)% following the implementation of decimal pricing in 2001 (see later section on decimal vs. fractional pricing).

Further, several papers show that it possible to infer the order processing, inventory and adverse selection components of the bid-ask spread from actual quote and trade data (the early literature includes Stoll, 1989; Huang and Stoll, 1997; Roll, 1984; Glosten and

Harris, 1988; Ho and Stoll, 1981; Madhavan and Smidt, 1991). As predicted by the theoretical analysis of the spread components, Lin et al. (1995) show that the order processing cost component exhibits economies of scale, hence it decreases with increases in trade size. Also, the adverse selection component increases with increases of trade size. Stoll (1989) shows that while spreads vary across stocks, the relative contribution of the components of the spread remain unchanged, hence the author estimates that the quoted spread consists of adverse selection costs (0.43%), order processing costs (0.47%) and inventory holding costs (0.10%). Boehmer (2005) shows that post-reform effective spreads for NYSE and NASDAQ stocks are $0.062 and $0.088 respectively. Also, the realised spread measures range from $0.035 for NASDAQ stocks to $0.011 for NYSE stocks. A comprehensive study of spread differences of exchanges around the world is presented by Jain (2003). The author shows that the NYSE has the lowest percentage quoted spread (0.20) amongst a sample of 51 exchanges. Spreads of emerging markets are considerably higher; Ukraine has the highest percentage quoted spread with 15.34%. A large proportion of cross-market spread variability is explained by market capitalisation and institutional design differences.

MARKET STRUCTURE AND TRADING COSTS

Tick size, discreteness and trading costs

One important question regarding trading costs is what would be the "true" quoted and effective spread if the tick size was absent? The tick size refers to the minimum permitted price variation on an exchange. On the NYSE, the current minimum tick size is $0.01, which was implemented on January 2001. The discrete nature of prices has long been considered as a major impediment to reducing transaction costs. Also, apart from the cost of trading in discrete prices, a second cost is added on trade prices when trades choose to trade in a set of prices that is smaller than the minimum set of prices allowed. This second feature of the trading process refers to price clustering and is usually associated with a reduction in negotiation costs, price uncertainty or round number preferences (see Harris, 1991, Ball et al., 1985 and Goodhart and Curcio, 1991). Harris (1999) predicts that a reduction in the minimum tick size will generally lead to a narrowing of the quoted and effective spread, a reduction of market depth (the number of shares that are available for each stock) and increase in price improvements as stepping ahead of other trades becomes cheaper under a smaller tick size.

The way in which price clustering and discreteness increase trading costs is presented in Hasbrouck (1999). In this study, the spread components are consecutively added on the "true" price. Thus, the permanent price (ask or bid) component follows a random walk, and the dealer's ask or bid price is the sum of the permanent price component plus a stochastic component that encompasses adverse selection, inventory and order processing costs. Translating these costs to quoted bids and asks entails rounding them up or down to the nearest discrete minimum tick. In a subsequent model, Ball and Chordia (2001) show that the "true" quoted spreads that would result in the market in the absence of minimum tick regulations vary from 11% to 24% of the "discrete" quoted spread. For example, the authors show that for a stock that is quoted at a spread of $0.20, the "true" spread varies from $0.02 to $0.06.

Spread Clustering and the Opportunity for Collusion in Dealer Markets

One important implication of the minimum tick rule was the assertion that NASDAQ market makers may have implicitly colluded in order to maintain wide spreads. Smidt (1971) rationalized the idea that a market with competing market makers would result in narrow spreads and that was the expectation for NASDAQ, a dealer market. Christie and Schultz (1994) compare spreads for a sample of 100 of the most actively traded firms listed on NASDAQ with a sample of firms listed on NYSE and the American Exchange (AMEX). At the time of the study, the minimum tick size was one-eighth of a dollar and Christie and Schultz (1994) show that for 70 of 100 NASDAQ firms the absence of odd-eights implies a minimum spread of two-eighths or $0.25. In contrast to NASDAQ firms, spreads of firms with similar trade characteristics listed on NYSE or AMEX are uniformly distributed across the full range of eighths. The allegation that NASDAQ market makers implicitly colluded to maintain high trading costs was further reinforced by the findings of Christie et al. (1994) that upon the publication of the findings of Christie and Schultz (1994) effective spreads in NASDAQ fell by approximately 50%. Christie and Schultz (1999) further show that market makers that use odd-eighths are much more likely to offer spreads of one-eighth hence bringing costs down. Bessembinder (1997) decompose spreads to order processing, adverse selection and inventory costs. The author reports that realized spreads, spreads net of adverse selection costs, on NASDAQ are positively related with price rounding practices of market makers. The same finding is not supported for a sample of same-size matched NYSE stocks that strengthens the collusion hypothesis for NASDAQ stocks.

The "collusion hypothesis" had overarching effects for trading cost and price clustering studies across the world (see Ahn et al., 2005 and Chiao and Wang, 2009). Dutta and Madhavan (1997) investigate the conditions under which market makers could have implicitly colluded, that is the conditions that would still result to spreads above competitive levels without however the explicit agreement of market makers (see also Kandel and Marx, 1997). The results have important implications for market design, supporting the argument that a smaller tick size should lead to more competitive spreads. Godek (1996) shows that preference trading, the regulation that allows market makers to direct orders to the market makers with the best trade price instead of the market maker with the best quote price, provides no incentives to market makers to offer competitive quote prices as almost all trades are preference trades in the sense that the terms of trade, including the trade price, are agreed prior to the trade. Thus, in this case, spreads are no longer an effective measure of trading costs and NASDAQ rules make sure that trading takes place at the best trade price. Hansch et al. (1999) show that preferenced trades on the London Stock Exchange (LSE) receive worse execution costs than non-preferenced trades. Also, internalized trades, trades that are submitted by a broker to a market maker of the same firm, receive better execution costs than non-internalized trades. Similarly, Bernhardt et al. (2005) show that execution costs are strongly influenced by the relationship between brokers and dealers, hence brokers that trade consecutively with the same dealer tend to receive greater price improvements and smaller execution costs. These findings support the hypotheses that there is a cost to negotiating quotes and also that broker-to-market maker relationships have important implications for trading costs. Simaan et al. (2003) show that market makers are more likely to offer narrow spreads when they can do so anonymously, emphasizing the benefits of increased competition for market maker quotation behaviour.

Market Structure Changes and Trading Costs

The effect of competition and regulation on trading costs

From 1997, the U.S. Securities Exchange Commission (SEC) allowed limit orders to compete with market maker quotes on NASDAQ. In particular, the "limit order display rule" that became effective in June 1997 required that limit orders that are better than the quotes submitted by market makers would be allowed on NASDAQ hence abandoning the monopoly of market makers to post quotes. The second SEC rule required market makers to publicise their best bid and ask quotes. Both actions intended to create greater market transparency and enhanced competition. In a similar natural experiment, Foerster and Karolyi (1998) show a fall in trading costs for firms that decide to list in multiple markets. Chung and Van Ness (2001) investigate the effect of the implementation of the order handling rules on trading costs. The authors show that trading costs fell substantially following the implementation of the rules. However, subsequent studies by Bessembinder (1999) and Chung et al. (2002) which compare post-reform trading costs for NASDAQ and NYSE stocks show that trading costs continue to be higher at NASDAQ than at the NYSE. In particular, Bessembinder (1999) reports that quoted spreads are 0.78 (1.03)% of share price on the NYSE (NASDAQ) and spreads are narrower on the NYSE for 77% of firms with equal trade characteristics.

Electronic vs. open outcry market structures

An important development in market microstructure has been the introduction of electronic trading which in most cases replaced the trading floors. The abandonment of the open outcry and the introduction of electronic trading has given rise to a series of studies that investigate to what extent trading costs were affected by this shift. Aitken et al. (2004) investigate the natural experiment of the introduction of electronic trading systems at the LSE, the Sydney Futures Exchange (SFE) and the Hong Kong Futures Exchange (HKFE). Aitken et al. (2004) show that trading costs fell significantly under the electronic trading system, a finding that is consistent for all three exchanges and holds even after controlling for volatility and trading volume effects. However, the authors also report that costs under the floor trading system were less prone to volatility spikes, hence under the electronic trading system, spreads tend to widen much faster in very volatile periods. Tse and Zabotina (2001) further show that trades submitted under an electronic trading system tend to be more susceptible to inventory cost considerations and tend to have a smaller information content than trades executed in the open outcry market (see also Huang, 2004). In the Foreign Exchange (FX) market, Ding and Hiltrop (2010) investigate the transfer of FX services from phone-based technology to electronic brokers systems. Ding and Hiltrop (2010) show that spreads narrowed after the introduction of electronic trading systems. However, spreads of informed traders increased, a finding that is related to an increase in market transparency. Finally, an excellent review of how the electronic trading has altered the U.S. stock markets is offered by Stoll (2006). The author decomposes costs to three components, commission costs, trading losses that reflect adverse selection costs and bid-ask spreads. Stoll (2006) notes that all three cost components have been reduced with the introduction of electronic trading as there has been a decrease in order handling costs, an increase in economies of scale and firm consolidations. The spread changes are significant even after controlling for the effect of tick size decreases and the introduction of stricter rules that enhanced competition and reduced the opportunities for excess market maker profits (see previous section on price clustering).

Decimal vs. fractional pricing

In January and April 2001 respectively, the NYSE and NASDAQ replaced the fractional pricing system with decimal pricing. This also implied a substantial reduction in tick size.

Several "before vs. after" studies show that post-decimalisation spreads have been dramatically reduced for both markets (see Bacidore et al., 2003 and NYSE, 2001). Bessembinder (2003) report that two important measures of trading costs exhibit significant improvement following decimalisation for the NYSE: effective spreads dropped substantially and price improvements, the tendency of stepping ahead of other quotes, significantly increased. For NASDAQ, effective and quoted spreads dropped, however price improvement rates did not improve as NASDAQ's rule allows traders to circumvent the price time priority rules by pre-arranging trades. Overall, Bessembinder (2003) shows that trade execution costs are quite similar for like-for-like stocks trading on NASDAQ and the NYSE. Further, ap Gwilym et al. (2005) show a widening in spreads, measured in ticks, following the introduction of decimal prices in the UK Long Gilt market. Such a result is anticipated if the new tick size is larger than the effective tick size. Nevertheless, ap Gwilym et al. (2005) show that the monetary value of spreads dropped significantly after the introduction of decimal prices which reflects an overall improvement in trading costs.

Trading Mechanisms and Trading Costs

Exchanges across the globe operate on a set of trading mechanisms that govern how orders are communicated and executed, the minimum tick regulations and the role of different market participants. In dealer markets, dealers "make the market" in the sense that investors buy and sell at the dealer's ask and bid price respectively. In an auction or limit order market, investors buy and sell from standing limit orders. In hybrid auction markets, liquidity is also facilitated by specialists or dealers. Regarding the differences between auction and dealer markets, several issues regarding market transparency and market liquidity arise, however the most important factor of market quality is execution costs. Several studies focus on this latter aspect of market quality vs. market structure. Christie and Huang (1994) study trading costs for firms that decided to move from NASDAQ, a dealer market, to NYSE or AMEX, two auction markets. The authors show that firms receive on average a net execution cost gain of 4.7 and 5.2 cents per share for the NYSE and AMEX respectively. These cost savings are persistent and generally attributed to liquidity enhancements for small stocks that are established on the NYSE and AMEX but not on NASDAQ. Huang and Stoll (1996) further decompose cost differences between NYSE and NASDAQ. They show that spread differences between the two markets cannot be attributed to differences in adverse selection costs, hence the realized spread components of stocks trading on NASDAQ and the NYSE are the same. Instead, Huang and Stoll (1996) argue that the absence of a limit order book and interdealer trading, such as internalisation and preferencing, are to be associated with the higher costs observed on NASDAQ. Heidle and Huang (2002) extend the sample of Christie and Huang (1994) by including stocks that decide to move from AMEX or the NYSE to NASDAQ. The authors show that firms moving to NASDAQ exhibit wider spreads (see also Clyde et al., 1997), most important however is the finding that the greater execution costs on NASDAQ are clearly associated with a higher probability of encountering informed traders on a dealer market setting. In this respect, trading costs and the implied probability of informed trading (as opposed to the explicit adverse selection component of the spread) are clearly related. Venkataraman (2001) compared matched samples of stocks trading on Paris Bourse, an automated limit order market with no specialists, and the NYSE. The results show that firms with similar trading characteristics tend to exhibit higher transaction costs on Paris Bourse even after controlling for adverse selection and economic cost differences. These findings show a clear distinction between a hybrid and a fully-automated order-book system. Most importantly, however, Venkataraman (2001) shows that hybrid markets are able to absorb large trades without a significant movement in price and also the human participation is more effective in reducing (transitory) volatility shocks.

An important question regarding trading costs is to what extent differences in market structure may be associated with differences in trading costs. Most of the studies that are concerned with market structure differences take one of the following two approaches: the "before vs. after" studies look at how trading costs change after the implementation of a new policy e.g. a decrease in the minimum tick or the implementation of new order handling rules (see Bessembinder, 1999 and Chung et al. 2002). The "A market vs. B market" studies focus on comparing stocks from two different exchanges, usually matched on trading characteristics (see Bessembinder, 1997 and Venkataraman, 2001). However, a third set of studies looks at differences in trading costs of dual-listed firms hence controlling for firm-specific characteristics. An early theoretical work in this area is conducted by Chowdhry and Nanda (1991). The authors model the informational efficiency of stocks that trade in multiple locations, showing that informed traders choose where to trade on the basis of fixed and informational costs. The latter finding implies that the market containing the larger pool of liquidity traders will dominate all other markets and will attract the largest portion of informed trading. Huang and Stoll (2001) consider a set of firms that trade on both the LSE and the NYSE. The authors show that the average quoted spread as a percent of quote midpoint is 1.09% for the LSE and 0.64% for the NYSE. Similarly, the effective spread as a percent of quote midpoint is 0.60% for the LSE and 0.46% for the NYSE. Consistent with the predictions of Harris (1999), Huang and Stoll (2001) show that the LSE exhibits greater depth.

Block trading

Block trades have received particular attention in the literature. This has been for two main reasons: first, block trades may potentially reveal private information and second the price impact of block trades is larger than the price impact of smaller, retail trades (see Seppi, 1990). Several studies show that institutional traders receive substantial discounts on commission costs (see Edmister, 1978, and Edmister and Subramaniam, 1982, Brennan et al., 1988). Markets with an off-order book trading facility (also known as upstairs markets) are the preferred trading venue for large trades. The increased trade transparency that an upstairs market offers protects dealers from adverse selection costs, hence reducing execution costs for block trades that otherwise would have to be traded on the order book. Bessembinder and Venkataraman (2004) study the upstairs market of the Paris Bourse. Their results show that on average block trades executed upstairs are executed with 20% of the total execution costs had these trades been executed on the limit order book. These discounts largely come from tapping into hidden liquidity pools, a feature which would have been impossible for an order book market. In a similar study for the Toronto Stock Exchange (TSE), Smith et al. (2001) show that large block trades receive discounts (price improvements) when executed upstairs and the information content of these trades is much lower than the information content of trades executed on the order book. Further, Smith et al. (2001) report that the upstairs market does not compete with the order book for liquidity. Instead, traders are encouraged to trade upstairs when there is insufficient liquidity and large spreads in the order book.

HIGH FREQUENCY TRADING (HFT) AND DIRECTIONS OF FUTURE RESEARCH

With only a brief examination of the current status of the literature on trading costs, one can easily identify that exchanges are currently shaken up by the growth of HFT firms and algorithm trading strategies. Iati (2009) approximated that HFT firms that account for approximately 2% of the total number of trading firms in the U.S., are responsible for almost 75% of trading on U.S. markets. HFT firms accounted for 55% of total trading volume in 2011, up from 26% in 2006 (Source: Bloomberg). A second, more specific definition of HFT refers to low-latency trading (see Habrouck, 2010). Hasbrouck (2010) defines latency as “the time it takes to observe a market event (e.g., a new bid price in the limit order book), through the time it takes to analyze this event and send an order to the exchange that responds to the event” Hasbrouck (2010, p.1). Under this definition, low latency strategies refer to strategies that are only detectable in the “millisecond environment”. Under these extreme trading conditions, trading firms have invested large amounts in technology that would allow them to gain execution priority even for a thousandth of a millisecond. Also, the exchanges have invested in technology that facilitates HFT as there is a general consensus that HFT replaces traditional market makers as providers of liquidity (see Menkveld, 2011).

Hasbrouck (2010) emphasizes that for HFT to be beneficial for the longer-term market investors, the effect of low latency trades will have to be observed by the latter. Hence, the author investigates the effect of low latency trades at 10-minute trade intervals which are easily observed by longer-term investors. Hasbrouck (2010) reports that HFT activity is key in reducing effective and quoted spreads and in decreasing short-term volatility, even for the down markets of 2008. Hendershott et al. (2011) ask a similar question for algorithm trades which resemble high frequency trades nevertheless they can execute at much higher latency than the latter. In general, Hendershott et al. (2011) investigate whether algorithm trades are mainly suppliers or demanders of liquidity. If algorithm trades are suppliers of liquidity, then they effectively help in reducing transaction costs as they operate alongside the traditional market makers. If however, algorithm trades are demanders of liquidity, then this may result in wider spreads, thereby increasing transaction costs. The authors show that algorithm trades have decreased transaction costs, especially for large cap stocks, a finding that is mainly attributed to a decrease in the adverse selection component of the spread. Hendershott and Riordan (2009) show that the decrease in the adverse selection component of the spread is associated with an increase in price discovery. The authors show that algorithm trades supply liquidity when markets are expensive and demand liquidity when markets are cheap, leading to greater price efficiency.

O'Hara et al. (2011) further formalize the consequences of HFT. In particular, the authors argue that in a high frequency world, clock time is of little importance for the needs of capturing informed trading and hence for controlling for the adverse selection transaction costs. This implies that market makers can no longer use order arrival rates to estimate the probability of informed trading as high frequency traders bet on tiny margins on a large number of trades which cancels the notion of clock time. O'Hara et al. (2011) refer to this problem as "flow toxicity", the tendency of market makers to provide liquidity at a loss. A key result from the classification of HFT as toxic trading, is the fact that market makers are unable to widen spreads when anticipating a greater liquidity cost. Easley et al. (2011) show that the "flash crash" of May 6th 2010 (which resulted in the biggest one-day point decline in the Dow Jones Industrial Average index history) could have been avoided had the market makers been able to calculate the increase in flow toxicity prior to the crash. The latest finding shows the clear relationship between trading cost and liquidity provision which however may collapse if insecurity (a large adverse selection cost component) prevails in the market as it forces market makers to exit the market.

SUMMARY AND CONCLUSIONS

This chapter examined bid-ask spreads, commissions and other trading costs. Recent evidence suggests that trading costs are at historically low levels, which paved the way for the implementation of algorithm trading which itself has led to substantial increases in trading volumes. Trading costs are classified as commission costs (which are determined by the exchange) and the cost components of the bid-ask spread (which are determined by market participants acting independently).

One of the main conclusions of this chapter is that bid-ask spreads can be decomposed to order processing, inventory and adverse selection costs which exist independently of each other and of commission costs. Evidence has shown that the 43% of the quoted spread consists of adverse selection costs, 47% of order processing costs and 10% of inventory holding costs (Stoll, 1989). It is shown that the variability in trading costs in exchanges around the world is a function of the variability in spread components. A recent study has shown that the NYSE has the lowest percentage quoted spread (0.20%) amongst a sample of 51 developed and emerging market exchanges. Trading costs for emerging markets appear to be considerably higher than for the more developed markets, with the highest being in Ukraine with average spreads of 15.34%. A second element of variability is attributed to differences in market structures. A large portion of this chapter is dedicated to the implication of market structure differences in trading costs. It is shown that electronic market trading has led to substantial decreases in both commission charges and bid-ask spreads, however, off-book trading is praised for the facilitation of large orders.

Finally, this chapter shows that future research will inevitably focus on the recent development of algorithm and high frequency trading. The dominance of high frequency trading has led to a deviation from the traditional trading cost measurement tools as high frequency traders thrive in low latency trading environments where clock time is of little importance. Current research on this has shown that high frequency trading tends to offer liquidity when liquidity is in short supply and demands liquidity when liquidity is abundant in the market, ultimately driving down spreads and enhancing the price discovery process. Nevertheless, the events of the May 6th 2010 Dow Jones crash are a clear example of the fragility of the markets to “order flow toxicity”, which derives the implementation of HFT techniques.

DISCUSSION QUESTIONS

1. What are the three components of the bid-ask spread?
2. Why are block trades important?
3. In high frequency trading, what is “order flow toxicity” and why is it important?
4. What are the differences between quoted, effective and realised spreads and price impact?
5. How have trading costs paved the way for the recent substantial increase in trading volume?

REFERENCES

Ahn, Hee-Joon, Jun Cai and Yan Leung Cheung. 2005."Price clustering on the limit-order book: Evidence from the Stock Exchange of Hong Kong." Journal of Financial Markets 8:4, 421-451.

Aitken, Michael J., Alex Frino, Amelia M. Hill and Elvis Jar. 2004. "The impact of electronic trading on bid-ask spreads: Evidence from futures markets in Hong Kong, London, and Sydney." Journal of Futures Markets 24:7, 675–696.

Amihud, Yakov and Mendelson, Haim. 1980. "Dealership Market: Market Making with Uncertainty." Journal of Financial Economics 8:1, 31-54.

ap Gwilym, Owain, Ian D. McManus and Stephen Thomas. 2005. "Fractional versus decimal pricing: Evidence from the UK long gilt futures market." Journal of Futures Markets 25:5, 419-442.

Bacidore, Jeffrey, Robert H. Battalio and Robert H. Jennings. 2003. "Order submission strategies, liquidity supply, and trading in pennies on the New York Stock Exchange." Journal of Financial Markets 6:3, 337-362.

Bagehot, Walter. 1971. "The Only Game in Town." Financial Analysts Journal 27:2, 12-14.

Ball, Clifford A. and Tarun Chordia. 2001. "True Spreads and Equilibrium Prices." Journal of Finance 56:5, 1801-1835.

Ball, Clifford A., Walter N. Torous and Adrian E. Tschoegl. 1985. "The degree of price resolution - the case of the gold market." Journal of Futures Markets 5:1, 29-43.

Bernhardt, Dan, Vladimir Dvoracek, Eric Hughson and Ingrid M. Werner. 2004. "Why Do Larger Orders Receive Discounts on the London Stock Exchange?" Review of Financial Studies 18:4, 1343-1368.

Bessembinder, Hendrik and Kumar Venkataraman. 2004. "Does an electronic stock exchange need an upstairs market?" Journal of Financial Economics 73:1, 3-36.

Bessembinder, Hendrik. 1997. "The degree of price resolution and equity trading costs." Journal of Financial Economics 45:1, 9–34.

Bessembinder, Hendrik. 1999. "Trade Execution Costs on NASDAQ and the NYSE: A Post-Reform Comparison." Journal of Financial and Quantitative Analysis 34:3, 387-407.

Bessembinder, Hendrik. 2003. "Trade Execution Costs and Market Quality after Decimalization." Journal of Financial and Quantitative Analysis 38:4, 747-777.

Boehmer, Ekkehart. 2005. "Dimensions of execution quality: Recent evidence for U.S. equity markets." Journal of Financial Economics 78:3, 463-704.

Brennan, Michael J., Tarun Chordia and Avanidhar Subrahmanyam. 1998. "Alternative factor specifications,security characteristics, and the cross-section of expected stock returns." Journal of Financial Economics 49:3, 345-373.

Chan, Kalok, Peter Y. Chung and Herb Johnson. 1995. "The Intraday Behavior of Bid-Ask Spreads for NYSE Stocks and CBOE Options." Journal of Financial and Quantitative Analysis 30:3, 329-346.

Chiao, Chaoshin and Zi-May Wang. 2009. "Price Clustering: Evidence Using Comprehensive Limit-Order Data." The Financial Review 44:1, 1-29.

Chowdhry, Bhagwan and Vikram Nanda. 1991. "Multimarket trading and market liquidity." Review of Financial Studies 4:3 483-511.

Christie, William G. and Paul H. Schultz. 1994. Why do NASDAQ market makers avoid odd-eighth quotes? Journal of Finance 49:5, 1813-1840.

Christie, William G, Harris, Jeffrey H., and Paul H. Schultz. 1994. "Why did NASDAQ market makers stop avoiding odd-eighth quotes?" Journal of Finance 49:5, 1841:1860.

Christie, William G. and Paul H. Schultz. 1999. "The initiation and withdrawal of odd-eighth quotes among Nasdaq stocks: an empirical analysis." Journal of Financial Economics 52:3, 409–442.

Chung, Kee H. and Robert A. Van Ness. 2001. "Order handling rules, tick size, and the intraday pattern of bid-ask spreads for Nasdaq stocks." Journal of Financial Markets 4:2, 143-161.

Chung, Kee H., Bonnie F. Van Ness and Robert A. Van Ness. 2002. "Spreads, Depths, and Quote Clustering on the NYSE and Nasdaq: Evidence after the 1997 Securities and Exchange Commission Rule Changes." Financial Review 37:4, 481–505.

Chung, Kee H., Charlie Charoenwong and David K. Ding. 2004."Penny pricing and the components of spread and depth changes." Journal of Banking & Finance 28:12, 2981-3007.

Clyde, Paul, Paul Schultz and Mir Zaman. 1997. "Trading Costs and Exchange Delisting: The Case of Firms that Voluntarily Move from the American Stock Exchange to the Nasdaq." Journal of Finance 52:5, 2103-2112.

Copeland, Thomas E. and Dan Galai. 1983. "Information Effects and the Bid-Ask Spread." Journal of Finance 38:5, 287-305.

Demsetz, Harold. 1968. "The cost of transacting." Quarterly Journal of Economics 82:1, 33-53.

Ding, Liang and Jonas Hiltrop. 2010. "The electronic trading systems and bid-ask spreads in the foreign exchange market." Journal of International Financial Markets, Institutions & Money 20:4, 323–345.

Dutta, Prajit K. and Madhavan, Ananth. 1997. "Competition and Collusion in Dealer Markets." Journal of Finance 52:1, 245-276.

Easley, David A., Marco M. Lopez de Prado and Maureen O'Hara. 2011. "The Microstructure of the 'Flash Crash': Flow Toxicity, Liquidity Crashes, and the Probability of Informed Trading." Journal of Portfolio Management 37:2, 118-128.

Edmister, Robert O. 1978. "Commission cost structure: Shifts and scale economies." Journal of Finance 33:2, 477-486.

Edmister, Robert O. and N. Subramanian. 1982. "Determinants of brokerage commission rates for institutional investors: A note." Journal of Finance 37:4, 1087-1093.

Foerster, Stephen R. and G. Andrew. Karolyi. 1998. "Multimarket trading and liquidity: a transaction data analysis of Canada-US interlistings." Journal of International Financial Markets, Institutions and Money 8:3-4, 393-412.

Garman, Mark. 1976. "Market Microstructure." Journal of Financial Economics 3:3, 257-275.

Glosten, Lawrence R. and Lawrence E. Harris. 1988. "Estimating the components of the bid/ask spread." Journal of Financial Economics 21:1, 123-142.

Glosten, Lawrence R. and Paul R. Milgrom. 1985. "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders". Journal of Financial Economics 14:1, 71-100.

Godek, Paul E. 1996. "Why Nasdaq market makers avoid odd-eighth quotes." Journal of Financial Economics 41:3, 465–474.

Goodhart, Charles. and Riccardo Curcio. 1991. "The clustering of bid/ask prices and the spread in the foreign exchange market." London School of Economics, Financial Market Group, Discussion Paper 110.

Hansch, Oliver, Narayan Y. Naik and S. Viswanathan. 1999. "Best Execution, Internalization, Preferencing and Dealer Profits." Journal of Finance 54:5, 1799-1828.

Harris, Lawrence. 1991. "Stock price clustering and discreteness." Review of Financial Studies, 4:3, 389-415.

Harris, Lawrence. 1999. "Trading in Pennies: A Survey of the Issues." Working Paper, University of Southern California.

Hasbrouck, Joel and Gideon Saar. 2010. "Low-latency trading." Working Paper. New York University.

Hasbrouck, Joel. 1999. "Security bid-ask dynamics with discreteness and clustering: Simple strategies for modelling and estimation." Journal of Financial Markets 2:1, 1-28.

Heidle, Hans G. and Roger D. Huang. 2002. "Information-Based Trading in Dealer and Auction Markets: An Analysis of Exchange Listings." Journal of Financial and Quantitative Analysis 37:3, 391-424.

Hendershott, Terrence, and Ryan Riordan. 2009. "Algorithmic Trading and Information." NET Institute Working Paper No. 09–08.

Hendershott, Terrence, Charles M. Jones, and Albert J. Menkveld. 2011. "Does Algorithmic Trading Improve Liquidity?" Journal of Finance 66:1–33.

Ho, Thomas and Hans R. Stoll. 1981. "Optimal Dealer Pricing under Transactions and Returns Uncertainty." Journal of Financial Economics 9:1, 47-73.

Ho, Thomas and Hans R. Stoll. 1983. "On Dealer Markets Under Competition", Journal of Finance 35:2, 259-267.

Huang, Roger D and Hans R. Stoll. 1997. "The Components of the Bid-Ask Spread: A General Approach." Review of Financial Studies 10:4, 995-1034.

Huang, Roger D. and Hans R. Stoll. 1996. "Dealer versus auction markets: A paired comparison of execution costs on NASDAQ and the NYSE." Journal of Financial Economics 41:3, 313-357.

Huang, Roger D. and Hans R. Stoll. 2001. "Tick Size, Bid-Ask Spreads, and Market Structure." Journal of Financial and Quantitative Analysis 36:4, 503-522.

Huang, Yu Chuan. 2004. "The components of bid-ask spread and their determinants: TAIFEX versus SGX-DT." Journal of Futures Markets 24:9, 835–860.

Iati, Robert. 2009. “High Frequency Trading Technology." TABB Group.

Jaffee, Jeff and Robert Winkler. 1976. "Optimal Speculation against an Efficient Market." Journal of Finance 31:1, 49-61.

Jain, Pankaj K. 2003. "Institutional Design and Liquidity at Stock Exchanges around the World." working paper, University of Memphis.

Jones, Charles M. 2002. "A century of stock market liquidity and trading costs." Working paper, Columbia University.

Jones, Charles M., and Marc L. Lipson. 2001. "Sixteenths: direct evidence on institutional execution costs." Journal of Financial Economics 59:2, 253-278.

Kandel, Eugene and Leslie M. Marx. 1997. "Nasdaq market structure and spread patterns." Journal of Financial Economics 45:1, 61-89.

Keim, Donald B. and Ananth Madhavan. 1997. "Transactions costs and investment style: an interexchange analysis of institutional equity trades." Journal of Financial Economics 46:3, 293-319.

Logue, Dennis. 1975. "Market-Making and the Assessment of Market Efficiency." Journal of Finance 30:1, 115-23.

Madhavan, Ananth and Seymour Smidt. 1991. "A Bayesian model of intraday specialist pricing." Journal of Financial Economics 30:1, 99-134.

Madhavan, Ananth. 2000. "Market microstructure." Journal of Financial Markets 3:3, 205-258.

Menkveld, Albert J. 2011. "High frequency trading and the new market makers." Working Paper. VU University Amsterdam, Tinbergen Institute, and Duisenberg School of Finance.

NYSE. 2001. "Comparing bid-ask spreads on the New York Stock Exchange and NASDAQ immediately following NASDAQ decimalisation." NYSE Research.

O'Hara, Maureen and George S. Oldfield. 1986. "The microeconomics of market making." Journal of Financial and Quantitative Analysis 21:04, 361-376.

O'Hara, Maureen, David A. Easley and Marcos M. Lopez de Prado. (forthcoming). "Flow Toxicity and Liquidity in a High Frequency World." Review of Financial Studies.

Roll, Richard. 1984. "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market." Journal of Finance 39:4, 1127-1139.

Seppi, Duane J. 1990. "Equilibrium Block Trading and Asymmetric Information." Journal of Finance 45:1, 73-94.

Simaan, Yusif, Daniel G. Weaver and David K. Whitcomb. 2003. "Market Maker Quotation Behavior and Pretrade Transparency." Journal of Finance 58:3, 1247–1268.

Smidt, Seymour. 1971. "Which Road to an Efficient Stock Market: Free Competition or Regulated Monopoly?" Financial Analysts Journal 27:5, 18-68.

Smith, Brian F. 2001. "Upstairs Market for Principal and Agency Trades: Analysis of Adverse Information and Price Effects." Journal of Finance 56:5, 1723-1746.

Stoll, Hans R. 1978. "The Supply of Dealer Services in Securities Markets." Journal of Finance 33:4, 1133-1151.

Stoll, Hans R. 1989. "Inferring the Components of the Bid-Ask Spread: Theory and Empirical Tests." Journal of Finance 44:1, 115-34.

Stoll, Hans R. 2006. "Electronic Trading in Stock Markets." The Journal of Economic Perspectives 20:1, 153-174.

Tse, Yiuman and Tatyana V. Zabotina. 2001. "Transaction Costs and Market Quality: Open Outcry Versus Electronic Trading." Journal of Futures Markets 21:8, 713–735.

Venkataraman, Kumar. 2001. "Automated versus Floor Trading: An Analysis of Execution Costs on the Paris and New York Exchanges." Journal of Finance 56:4, 1445-1485.

William G. Christie and Roger D. Huang. 1994. "Market Structures and Liquidity: A Transactions Data Study of Exchange Listings." Journal of Financial Intermediation 3:3, 300-326.

ABOUT THE AUTHOR

Thanos Verousis is a Lecturer in Finance at Bangor Business School, Bangor University, UK. Thanos' research interests broadly lie in the microstructure of financial markets, however his main contribution is on the microstructure of individual equity options trading at NYSE LIFFE. He has also conducted research in the area of micro analysis of agent behaviour in financial markets and in the liquidity aspects of investment trusts. His work has appeared at the Journal of Futures Markets, the European Journal of Finance and the International Review of Financial Analysis. Dr Verousis has also taught at Swansea University and holds a PhD in Financial Markets from the University of Wales.

ACKNOWLEDGEMENTS

The author would like to thank Timotheos Angelidis and Owain ap Gwilym for their valuable support and advice.

CHAPTER 20 BID-ASK SPREADS, COMMISSIONS AND OTHER COSTS

1. What are the three components of the bid-ask spread?

   The order processing cost component is referred to as the "price of immediacy" and traces its origins to the problem of simultaneity in buying and selling securities. A trader wishing to buy (sell) an asset has no guarantee that a seller (buyer) will be readily available in the market to provide this service, hence traders will only agree to provide this service if they are compensated. The bid-ask spread reflects the premium for providing this service. The inventory cost component arises from short-term order imbalances as dealers adjust their spreads in order to maximize profits and fulfil their requirement as providers of liquidity. The adverse selection component of the spread arises as dealers widen their spread in order to be compensated from trading with informed traders.

2. Why are block trades important? How do exchanges deal with block trades?

   Block trades are important because of the potential price impact that they carry and because they tend to be more informative than the smaller, retail trades. In the US, untill 1974, the commission charge on block trades was relative to commission charges to smaller trades, however, after 1974, block trades started receiving large discounts. Exchanges facilitate large orders via off-book trading floors that reduce execution costs and increase trade transparency. These discounts largely come from tapping into hidden liquidity pools, a feature which would have been impossible had the block trade been executed in the order book market.

3. In high frequency trading, what is "order flow toxicity"?

   The term "order flow toxicity" refers to the tendency of market makers to provide liquidity at a loss. The problem of "order flow toxicity" arises as a consequence of high frequency trading where market makers can no longer use order arrival rates to estimate the probability of informed trading as high frequency trades bet on tiny margins on a large number of trades which cancels the notion of clock time. In market microstructure, “order flow toxicity” is important as the presence of “toxic” orders may potentially drive market makers out of the market. The “flash crash” of May 6th 2010 is a good example of the potential consequences of “order flow toxicity”.

4. What are the differences between quoted, effective and realised spreads and price impact?

The simplest measure of implied execution costs is the quoted spread, which is usually denoted in percentage basis points. The quoted spread of a stock is calculated as the difference between ask and bid prices at a given point in time divided by the quote midpoint (which is the average of the ask and bid price at the same time). Quoted spreads are only implicit measures of execution costs, as they do not refer to the actual traded price. Effective bid-ask spreads are estimated as the absolute percentage difference between the traded price and the quote midpoint. In markets where trade negotiations are allowed, the effective spread also reflects trade improvements as traders are permitted to trade “inside” the quotes, hence the effective spread tends to be lower than the quoted spread. Price impacts measure adverse selection costs, that is the costs of trading with an informed trader and are estimated as the percentage difference between the midquote that prevailed at the time of the trade and a future midquote. The realised spread is estimated as the difference between the effective spread and price impact. As the realised spread is net of the price impact, it reflects trading costs net of any losses to informed traders.

5. How have changes in trading costs paved the way for the recent substantial increase in trading volume?

First, technological advances with the increased use of electronic trading have substantially reduced the cost of handling orders by broker-dealers. Studies have shown that the average round-trip commission has fallen substantially over the recent years.

Second, the implementation of high frequency trading strategies relies on heavy volume to extract marginal profits from a very large number of trades. In 2009, HFT firms accounted for approximately 2% of the total number of trading firms, however they were responsible for almost 75% of the total trading volume in the U.S., approximately a threefold increase from 2006. The decrease in trading costs has made the implementation of these strategies possible.